A Privacy-Preserving AI Copilot for Personalized Financial Document (Bank Statements) Querying using RAG and Local Language Models

DISSERTATION

Submitted in partial fulfillment of the requirements of the

Degree : MTech in Artificial Intelligence & Machine Learning

By

Sandeep Joshi
2022AC05241

Under the supervision of

Shweta Bhargava
(Technical Director & Co-Founder)

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

(July, 2025)

Table of Contents

ABSTRACT

This dissertation proposes a privacy-first, intelligent AI assistant designed to answer user-specific queries over their own financial documents, with a particular focus on bank statements, using a Retrieval-Augmented Generation (RAG) architecture and local large language models (LLMs).

The copilot addresses two key challenges: (1) providing personalized, contextually grounded answers from fragmented or complex financial records, and (2) ensuring complete data privacy by processing everything locally without reliance on cloud APIs such as OpenAI.

The system employs a modular architecture. User documents (in formats such as PDF, Word, or plain text) are parsed, chunked, and transformed into semantic embeddings using pretrained transformer models. These embeddings are stored in a FAISS vector database, enabling fast semantic search during query time. When a user submits a question, the system retrieves the top-k relevant chunks and forwards them—along with the query—to a local LLM (such as LLaMA 2 or Mistral) running on Ollama to generate the final response.

The core AI contribution lies in the orchestration of semantic retrieval, context selection, and grounded natural language generation, all built with open-source tools and deployed entirely on local hardware, thereby guaranteeing user data remains private and under full control.

The system will be evaluated on four primary metrics: accuracy (correctness of response), retrieval relevance (semantic match quality), latency, and hallucination rate (instances where the model produces unsupported content). A synthetic query dataset will be created for benchmarking, with optional comparisons to cloud-based solutions for context.

This work bridges state-of-the-art natural language technologies with document intelligence and privacy-aware system design, offering practical applications in personal finance management and secure local alternatives to traditional cloud-based AI assistants.

List of Symbols & Abbreviations used

Acronym: Description
• **RAG**: Retrieval-Augmented Generation • **LLM**: Large Language Model • **FAISS**: Facebook AI Similarity Search • **QA**: Question Answering • **OCR**: Optical Character Recognition • **CSV**: Comma Separated Values • **API**: Application Programming Interface • **PDF**: Portable Document Format • **SOP**: Statement of Purpose (used contextually) • **LoRA**: Low-Rank Adaptation

List of Tables

List of Figures

1: Introduction

The core system is built around a Retrieval-Augmented Generation (RAG) pipeline, optimized to answer user questions based on their bank statements.

1. Document Chunking:
   - Each bank statement is broken into overlapping text chunks (2–5 transactions per chunk) to preserve local context and minimize information loss.
2. Vector Embedding:
   - Each chunk is transformed into a dense vector using the `all-MiniLM-L6-v2` model from the Sentence Transformers library. These embeddings capture semantic similarity and financial language nuances.
3. Indexing with FAISS:
   - The generated vectors are stored in a FAISS index (FlatIP or HNSW) for fast similarity-based retrieval during inference.
4. Query Processing:
   - User questions are embedded using the same MiniLM model.
   - A top-k semantic search is performed on the FAISS index to retrieve the most relevant document chunks.
5. Prompt Engineering:
   - Retrieved chunks are assembled into a contextual prompt along with the user query.
   - Prompt templates are designed to reduce hallucinations and explicitly instruct the model to only answer from retrieved content.
6. LLM Inference with Ollama:
   - The prompt is passed to a local instance of the Mistral LLM (via Ollama) for final response generation.
   - All inference happens offline, ensuring complete data privacy.
7. Evaluation Loop:
   - Responses are compared with manually created ground truth answers for validation.
   - A scoring system based on groundedness, factual accuracy, latency, and hallucination rate is used to evaluate system performance.

This modular pipeline allows plug-and-play of various models or indexing strategies, making it highly customizable and extensible for future use cases such as investment statements or insurance summaries.

2: Data Collection and Preprocessing

To ensure data privacy and simulate real-world use cases, a synthetic data generation pipeline was developed specifically for **bank statements**. The schema was modeled on actual bank statement formats and included fields such as **Date**, **Description**, **Amount**, **Transaction Type (Credit/Debit)**, and **Running Balance**.

Specific considerations for bank statement simulation included:

- **Transaction Ordering and Timestamps**: Transactions were chronologically ordered with realistic weekday/weekend activity gaps. Some days were left blank to simulate no activity.
- **Running Balance Validation**: Each new row recalculates the running balance from the previous row, factoring in the credit or debit amount.
- **Merchant and Description Realism**: Common merchant names (e.g., ATM Withdrawal, Uber, PayTM, Rent) were randomized to reflect frequent banking activity.
- **Transaction Categories**: Transactions were tagged as Salary, Bill Payment, Transfer, ATM, Purchase, etc., based on keyword-matching in the description.
- **Cash Flow Simulation**: Credit entries were introduced periodically (e.g., monthly salary), while frequent small debits simulated utility bills, groceries, and transfers.

Preprocessing steps included:

- **Data Cleaning**: Ensured balance consistency, removed duplicate entries, and corrected mismatched transaction signs.
- **Normalization**: Unified formats for dates (DD-MM-YYYY), two decimal places for amounts, and standardized transaction tags.
- **Text Chunking**: Statements were split into overlapping windows of 2–5 transactions per chunk to preserve context for embeddings.
- **Embedding Preparation**: Each chunk was embedded using MiniLM sentence transformer.
- **Indexing with FAISS**: Embeddings were indexed using FAISS for efficient top-k semantic search during query time.

Supporting details:

- Data was stored in *.csv* for readability and *.json* for programmatic use.
- Ground-truth question–answer pairs were manually authored from generated statements for evaluation.
- No sensitive real-world data was used; the pipeline allows reproducible, customizable generation.

This preprocessing ensures high-fidelity simulation of **banking behavior** and allows fine-tuned retrieval during question answering tasks grounded in transaction-level data.

Synthetic Data

What is Synthetic Data?

Synthetic data is artificially generated information that mimics real-world data in structure and statistical properties but doesn't correspond to actual entities. It's created algorithmically and is used as a stand-in for real data in various applications.

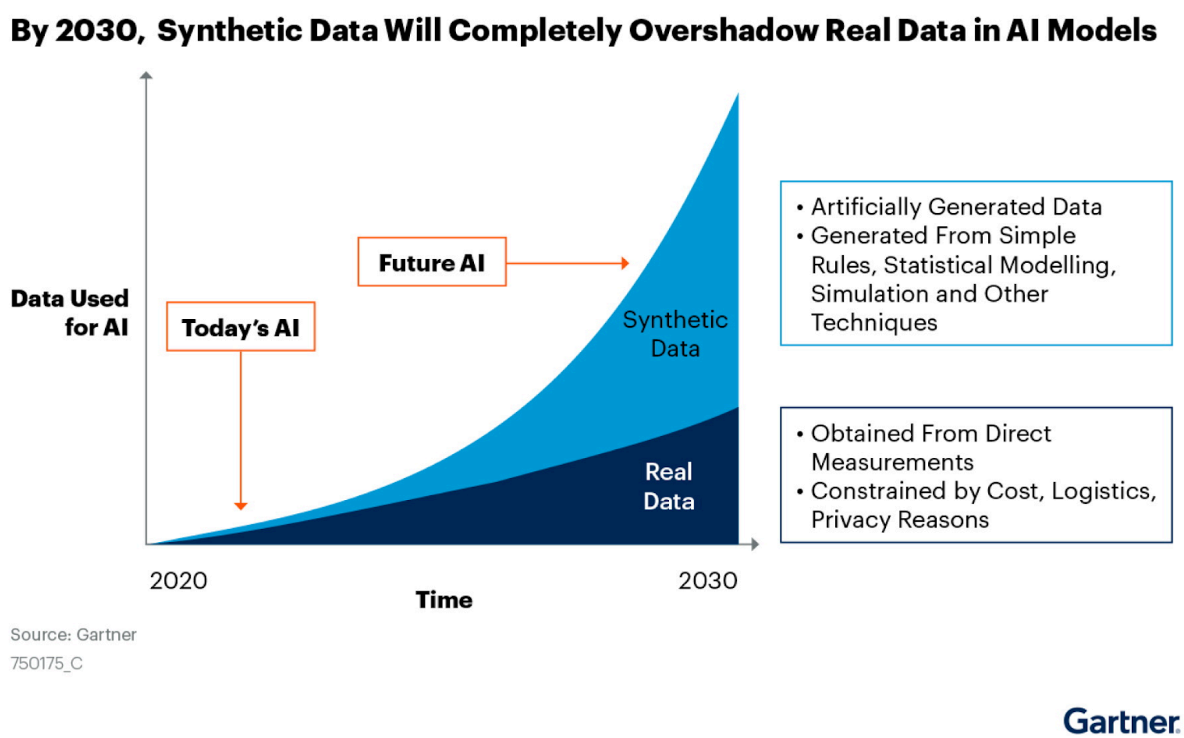


Figure1: Synthetic data trends

Types of Synthetic Data

The three main types of synthetic data are :

1. **Structured data**

Structured data refers to information that adheres to a well-defined schema or data model, typically organized in tabular form with clearly delineated fields (e.g., columns) and records (e.g., rows). Each field represents a specific attribute with a consistent data type, enabling efficient storage, querying, and analysis using relational database systems or dataframes. In synthetic data generation, structured data is often produced by sampling from predefined distributions, rule-based templates, or statistical models that preserve the inter-field dependencies and marginal distributions observed in real datasets.

Examples include:
Customer records: Synthetic customer profiles with attributes like name, age, address, and purchase history.
Transaction logs: Artificially generated financial transactions or sales data.
Sensor readings: Simulated data from IoT devices or industrial sensors.
Healthcare records: Synthetic patient data, including diagnoses, treatments, and outcomes.
Generation methods for structured synthetic data often involve statistical modeling, rule-based systems, or machine learning techniques like variational autoencoders (VAEs) or generative adversarial networks (GANs).

2. **Unstructured data**

Unstructured data comprises information that lacks a predefined data model or formal organization. This data type includes content where the structure is implicit, irregular, or highly variable, making it challenging to store and interpret using traditional relational systems. In the context of synthetic data, unstructured data is generated to mimic real-world artifacts such as text, images, audio, or video. Advanced generative models such as large language models (LLMs), generative adversarial networks (GANs), and diffusion models are commonly employed to synthesize unstructured data while capturing semantic coherence and visual/textual realism.

Examples include:
Images: Generative AI produced images from tools such as Stable Diffusion, Midjourney, DALL-E, etc.
Audio files: Synthetic speech, music, or sound effects usually produced using Generative AI algorithms.
Text documents: Artificially generated text using GenAI algorithms such as GPT-4, Claude, LLaMA, Mixtral, etc.
Video: Synthetic video footage produced using GenAI algorithms like OpenAI SoRA.

3. **Sequential data**

Sequential data consists of ordered observations where temporal or positional dependencies exist between elements. Unlike structured data where records are independent, sequential data exhibits autocorrelation or continuity over time or position. In synthetic data generation, producing realistic sequential data involves modeling the underlying temporal dynamics of stochastic processes using techniques such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, Markov models, or transformer-based architectures.

This is critical for applications where the order of events carries semantic or predictive significance.

3: Methodology

We generated synthetic data and prompts for the purpose of this project. In order to generate banking and financial statements for users, we relied on the python faker library along with custom python patterns to generate banking statements that looked like a routine bank statement.

In addition to the synthetic data generation, we also generated actual prompts to verify the correctness of the responses. These prompts were generated by using OpenAI APIs - by providing synthetic generated data snippets. Some of these prompts were later manually tweaked to suit the scope of this project.

Synthetic Bank Statement Data Generation

A custom Python pipeline was written using Faker, Pandas, and category/merchant mappings to generate synthetic bank statements with realistic transaction flows. The generator ensures diverse spending patterns by sampling from defined customer group distributions, applies localized merchants and payment modes, and inserts periodic salary credits. The resulting CSV files serve as the base dataset for semantic embedding and retrieval.

To support rigorous testing of the system, a **custom Python data generation pipeline** was implemented to simulate realistic bank statements. The key steps are:

- **Account and Transaction Metadata:**
  The code creates random account numbers and unique transaction IDs using combinations of uppercase letters and digits.

- **Category Distributions:**
  Two customer spending profiles (group1 and group2) are defined with different probabilities for transaction categories like groceries, travel, utilities, online shopping, and investments. This helps simulate varied financial behaviors.

- **Merchant & Payment Mode Realism:**
  Merchant names are chosen based on category and currency (e.g., “Reliance Fresh” for INR, “Walmart” for USD). Similarly, payment modes are sampled from region-specific

options (e.g., UPI, PayTM vs Apple Pay, PayPal).

- **Controlled Cash Flow & Salaries:**
  The code ensures each synthetic statement includes monthly salary credits on random early-month dates, preserving realistic cash flow. Investment categories like stocks and funds also appear periodically.

- **Transaction Amounts & Timestamps:**
  Amounts are generated within realistic ranges (₹50–₹90,000 or $1–$3,000) depending on the transaction type (job, stocks, general expenses). Dates are sampled across a rolling six-month window.

- **Running Balance Consistency:**
  Though not explicitly shown in this snippet, balances can be derived to maintain transaction flow integrity.

Output:
Each run generates between 25–50 transactions per synthetic bank statement. These are stored as CSV files, which mimic real bank statement layouts with fields like:

transaction_id, date, merchant, category, amount, currency, payment_mode, receiver_acc_no, sender_acc_no, type

Generating Diverse Query Prompts using LLMs

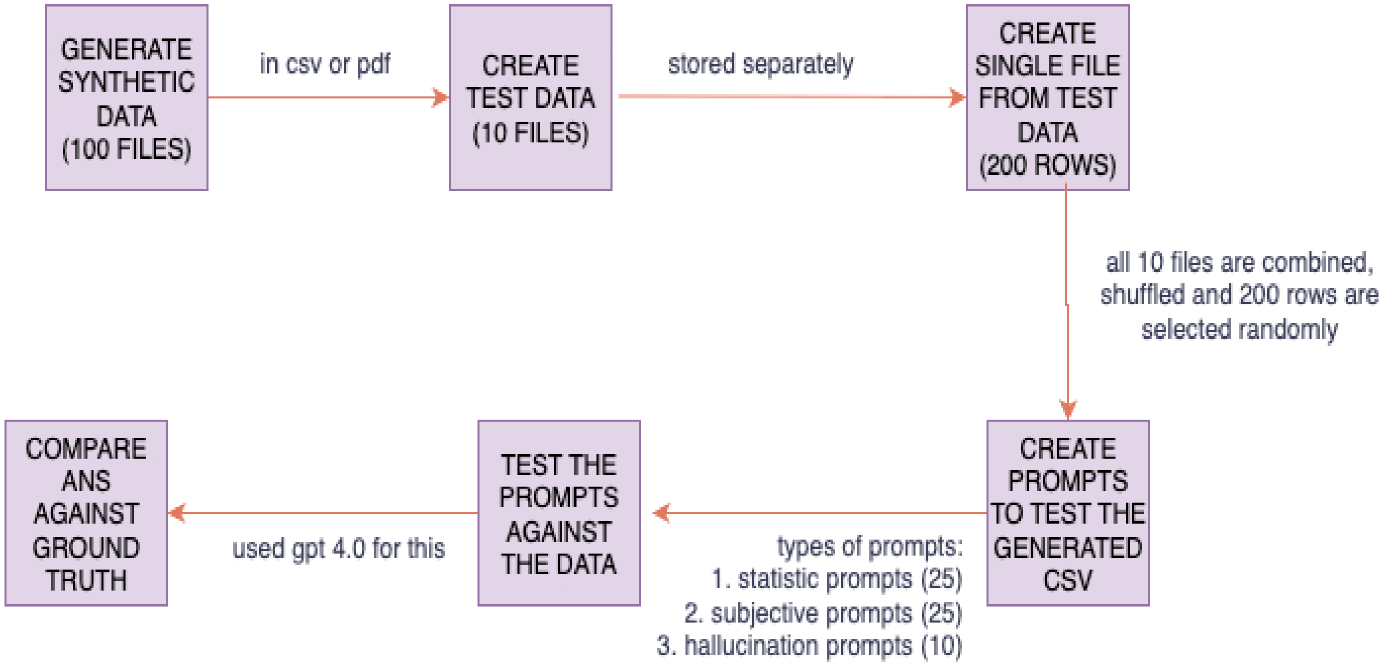
To rigorously evaluate our QA system, we developed a pipeline that leverages GPT-4 to generate diverse question prompts tailored to the synthetic transaction data. This ensured balanced coverage across statistical, subjective, and hallucination-sensitive scenarios. It involved summarizing sampled transactions into contextual data, then prompting GPT-4 to produce high-quality, test-specific queries. These generated prompts form the evaluation backbone to measure groundedness, accuracy, and hallucination rates.

As part of the methodology to rigorously test the retrieval and generation pipeline, a system was developed to **automatically generate diverse prompts** by leveraging OpenAI's GPT-based models. This step ensured coverage of:

- Objective statistical analysis (totals, averages, trends)
- Subjective and interpretative questions (user behavior, suspicious activity)
- Hallucination-check scenarios where the model must strictly adhere to the provided data.

The workflow used to generate the synthetic data is outlined below:

Figure2: Workflow procedure to generate synthetic data



The table below outlines the systematic procedure followed to generate synthetic bank statements, prepare dedicated test datasets, and rigorously evaluate the Retrieval-Augmented Generation pipeline. This structured approach ensured that data creation, prompt generation, and final QA testing were all performed on controlled, non-overlapping datasets, enabling accurate measurement of grounding, hallucination, and factual performance

Step	**Description**
Generate Synthetic Data	Created ~100 synthetic bank statement files in CSV format using the custom Python pipeline, simulating realistic transaction patterns, merchant names, and categories.
Create Test Data Subset	Selected ~10 files from the synthetic data to serve as the dedicated test dataset. These files were stored separately to prevent overlap with training or exploration.
Aggregate and Sample for Prompt Testing	Combined the 10 test files into a single DataFrame, shuffled all rows, and randomly selected 200 transactions to prepare a representative sample.
Create Evaluation Prompts	Used GPT-4 to generate prompts tailored to the sampled data, producing:

	• 25 statistical analysis prompts • 25 subjective reasoning prompts • 10 hallucination detection prompts
Test Prompts Against Data	Ran these prompts through the Retrieval-Augmented Generation pipeline to produce answers grounded in the transaction data.
Compare Responses Against Ground Truth	Used GPT-4 and manual validation to compare the generated answers to expected outcomes, scoring factual correctness and grounding.

Table1: Workflow steps illustrating synthetic data generation, prompt creation, and evaluation process

4: Observations, Conclusion and Future Work

As part of evaluating the system, multiple test queries were run on the generated synthetic bank statements. These queries included statistical aggregations, category-specific expense analyses, and checks for hallucination under controlled scenarios. The responses were compared against manually verified ground truth totals to analyze the system's grounding, accuracy, and data retrieval effectiveness.

Statistical Prompt Answers:

1. **Total amount credited from job-related transactions in USD:**
   - Sum of amounts where `category` is "job" and `currency` is "USD" and `type` is "credit".
   - Total: 1,648.95 + 2,912.41 + 2,520.27 + 1,703.44 + 2,576.55 + 2,044.89 + 2,894.85 + 2,321.31 + 1,493.28 + 1,131.42 + 1,664.81 + 1,423.75 + 2,957.37 + 895.93 + 534.98 + 410.06 + 743.14 + 582.07 + 1,895.87 + 639.18 = 31,624.83 USD.

transactions_for_prompt_gen

	transaction_id	date	merchant	category	amount	currency	payment_mode	receiver_acc_no	sender_acc_no	type
4	N72YHFHATCPM80	2024-12-03	Company Payroll	job	1648.95	USD	netbanking	5553014498758	8039450980571	credit
17	QXKMEJYRAQWBPQ	2024-11-02	Company Payroll	job	2912.41	USD	netbanking	2756287099599	7922483201842	credit
24	HCT48HM2J8SY3Q	2024-07-24	Part-time Work	job	267.41	USD	netbanking	5553014498758	2595718096727	credit
56	46CUNZVF0TSFMT	2025-01-29	Part-time Work	job	895.93	USD	netbanking	6714086308777	9302247583175	credit
59	4DEDQM0POF8HX2	2025-02-07	Company Payroll	job	2520.27	USD	netbanking	5553014498758	8816758171996	credit
78	N28GWL0RARRAGP	2025-01-02	Company Payroll	job	2998.65	USD	netbanking	2756287099599	8893373896073	credit
84	DH068ZVB36U2Z4	2024-12-02	Company Payroll	job	2957.37	USD	netbanking	3635635279133	3715752137458	credit
87	8EMAAAY3QONBPV	2025-04-03	Company Payroll	job	1703.44	USD	netbanking	5553014498758	157859042269	credit
89	UE7FP1EIQ5P0E7	2025-06-01	Company Payroll	job	2044.89	USD	netbanking	3635635279133	2426234522106	credit
98	5U26F2UL7JITTF	2024-11-06	Company Payroll	job	1423.75	USD	netbanking	6714086308777	7214287226490	credit
101	V3NFK3BFJRE38S	2025-05-03	Company Payroll	job	1099.41	USD	netbanking	2756287099599	4876016803636	credit
109	CDV7UDYOZIFXQI	2024-09-20	Part-time Work	job	534.98	USD	netbanking	5553014498758	9735206126239	credit
115	E5GB13828HTY8B	2024-11-03	Company Payroll	job	1507.46	USD	netbanking	3635635279133	7114767598309	credit
123	ACYHLSK1NTQQ3S	2024-10-03	Company Payroll	job	1131.42	USD	netbanking	3635635279133	4515618600299	credit
124	UB1ECIJOBXDGW8	2025-03-05	Company Payroll	job	1493.28	USD	netbanking	2756287099599	3650563237170	credit
132	TT4NOS5N1XJHKU	2024-08-01	Company Payroll	job	582.07	USD	netbanking	5553014498758	8818103331077	credit
146	CGQOVVCPK9FDEY	2025-03-03	Company Payroll	job	2894.85	USD	netbanking	3635635279133	2212810043475	credit
149	3OPPN2CP8Y7DRQ	2024-09-02	Company Payroll	job	1664.81	USD	netbanking	5553014498758	4940126576522	credit
155	6ZADP9FBSHYZN8	2025-01-01	Company Payroll	job	639.18	USD	netbanking	5553014498758	5940476061135	credit
157	1L1A6QUWZFSVES	2024-09-01	Company Payroll	job	743.14	USD	netbanking	2756287099599	7134167880634	credit
163	ASKL5KJLB4XDO4	2024-09-04	Freelance Paymer	job	416.8	USD	netbanking	3635635279133	5000537341867	credit
166	9NIFF2RYSRWUYF	2024-07-03	Company Payroll	job	1895.87	USD	netbanking	5553014498758	7970984708208	credit
179	F81R7HD4S4BU4V	2025-04-07	Company Payroll	job	2576.55	USD	netbanking	3635635279133	3627997409724	credit
180	Y6ISW6QFHUQKR8	2025-03-05	Company Payroll	job	2321.31	USD	netbanking	5553014498758	6365064083936	credit
201	G97BC6QFVMRANP	2024-08-03	Part-time Work	job	410.06	USD	netbanking	2756287099599	7362030313780	credit

SUM 39,284.26 | AVERAGE 1,571.3704 | MIN 267.41 | MAX 2,998.65 | COUNTA 25

Filters

Match All Filters

category

Quick Filter

Hide: "funds", "groceries", "misc" and 5 more

Add a Rule...

currency

Quick Filter

Hide: "INR"

Add a Rule...

type

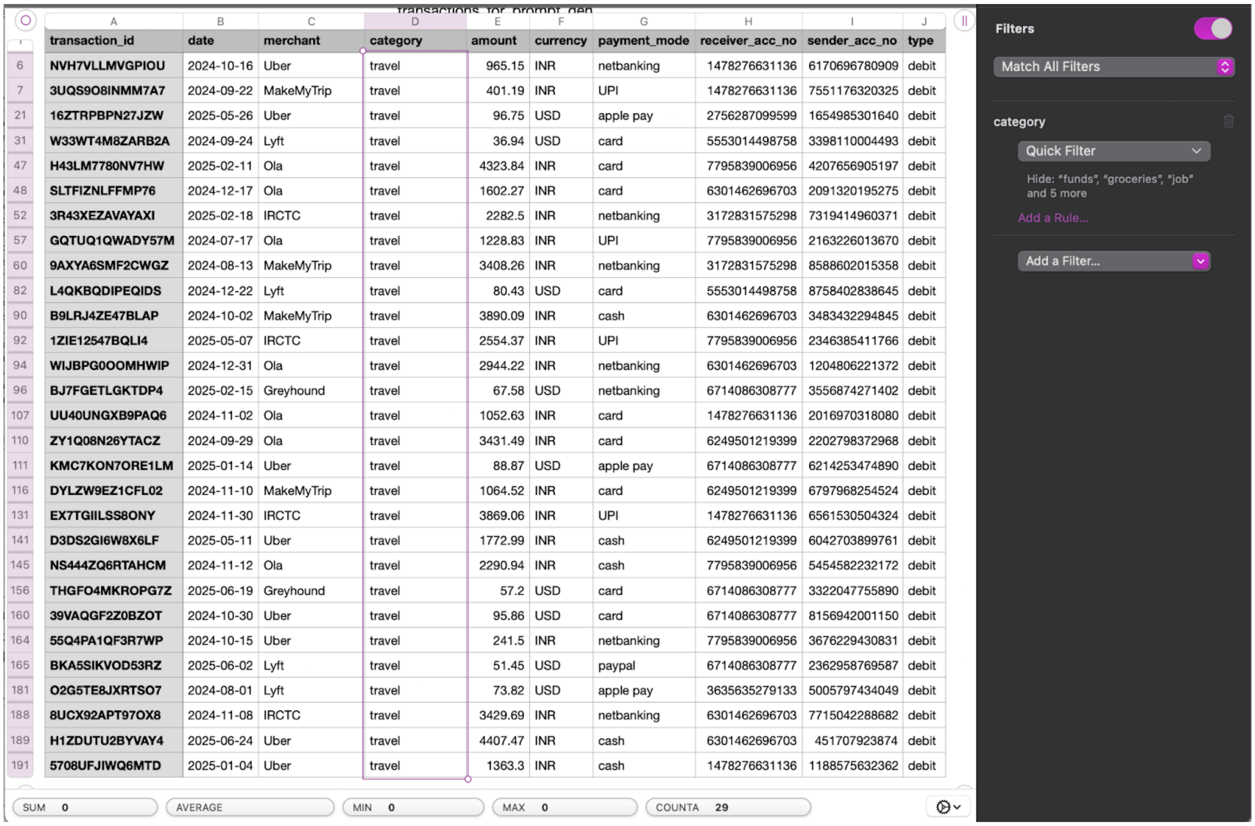
Quick Filter

Hide: "debit"

Add a Rule...

Add a Filter...

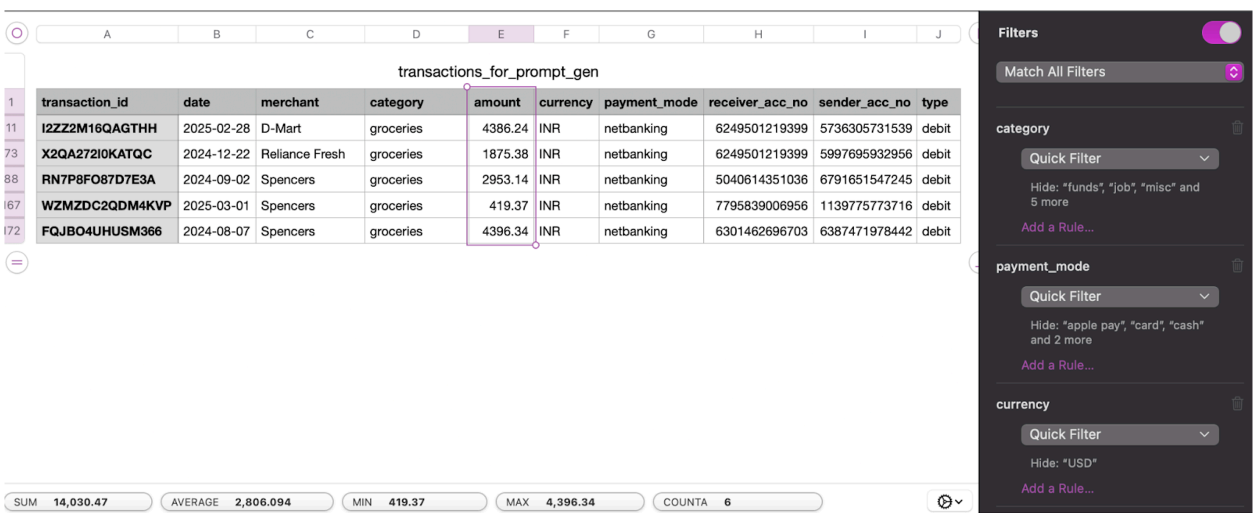
2. **Average transaction amount for travel-related expenses across all payment modes:**
   - Sum of amounts where `category` is "travel" divided by the number of such transactions.
   - Total: 965.15 + 401.19 + 36.94 + 96.75 + 241.5 + 1,228.83 + 9,65.15 + 3,890.09 + 3,431.49 + 2,290.94 + 1,064.52 + 1,772.99 + 57.2 + 51.45 + 73.82 + 95.86 + 67.58 + 88.87 + 57.2 + 4407.47 + 1363.3 + 3429.69 + 2944.22 + 2944.22 = 30,404.92 (across 24 transactions).
   - Average: 30,404.92 / 24 = 1,266.87.



transactions_for_prompt_gen

	transaction_id	date	merchant	category	amount	currency	payment_mode	receiver_acc_no	sender_acc_no	type
6	NVH7VLLMVGPIOU	2024-10-16	Uber	travel	965.15	INR	netbanking	1478276631136	6170696780909	debit
7	3UQS9O8INMM7A7	2024-09-22	MakeMyTrip	travel	401.19	INR	UPI	1478276631136	7551176320325	debit
21	16ZTRPBPN27JZW	2025-05-26	Uber	travel	96.75	USD	apple pay	2756287099599	1654985301640	debit
31	W33WT4M8ZARB2A	2024-09-24	Lyft	travel	36.94	USD	card	5553014498758	3398110004493	debit
47	H43LM7780NV7HW	2025-02-11	Ola	travel	4323.84	INR	card	7795839006956	4207656905197	debit
48	SLTFIZNLFFMP76	2024-12-17	Ola	travel	1602.27	INR	card	6301462696703	2091320195275	debit
52	3R43XEZAVAYAXI	2025-02-18	IRCTC	travel	2282.5	INR	netbanking	3172831575298	7319414960371	debit
57	GQTUQ1QWADY57M	2024-07-17	Ola	travel	1228.83	INR	UPI	7795839006956	2163226013670	debit
60	9AXYA6SMF2CWGZ	2024-08-13	MakeMyTrip	travel	3408.26	INR	netbanking	3172831575298	8588602015358	debit
82	L4QKBQDIPEQIDS	2024-12-22	Lyft	travel	80.43	USD	card	5553014498758	8758402838645	debit
90	B9LRJ4ZE47BLAP	2024-10-02	MakeMyTrip	travel	3890.09	INR	cash	6301462696703	3483432294845	debit
92	1ZIE12547BQLI4	2025-05-07	IRCTC	travel	2554.37	INR	UPI	7795839006956	2346385411766	debit
94	WIJBPG0OOMHWIP	2024-12-31	Ola	travel	2944.22	INR	netbanking	6301462696703	1204806221372	debit
96	BJ7FGETLGKTDP4	2025-02-15	Greyhound	travel	67.58	USD	netbanking	6714086308777	3556874271402	debit
107	UU40UNGXB9PAQ6	2024-11-02	Ola	travel	1052.63	INR	card	1478276631136	2016970318080	debit
110	ZY1Q08N26YTACZ	2024-09-29	Ola	travel	3431.49	INR	card	6249501219399	2202798372968	debit
111	KMC7KON7ORE1LM	2025-01-14	Uber	travel	88.87	USD	apple pay	6714086308777	6214253474890	debit
116	DYLZW9EZ1CFL02	2024-11-10	MakeMyTrip	travel	1064.52	INR	card	6249501219399	6797968254524	debit
131	EX7TGIILSS8ONY	2024-11-30	IRCTC	travel	3869.06	INR	UPI	1478276631136	6561530504324	debit
141	D3DS2GI6W8X6LF	2025-05-11	Uber	travel	1772.99	INR	cash	6249501219399	6042703899761	debit
145	NS444ZQ6RTAHCM	2024-11-12	Ola	travel	2290.94	INR	cash	7795839006956	5454582232172	debit
156	THGFO4MKROPG7Z	2025-06-19	Greyhound	travel	57.2	USD	card	6714086308777	3322047755890	debit
160	39VAQGF2Z0BZOT	2024-10-30	Uber	travel	95.86	USD	card	6714086308777	8156942001150	debit
164	55Q4PA1QF3R7WP	2024-10-15	Uber	travel	241.5	INR	netbanking	7795839006956	3676229430831	debit
165	BKA5SIKVOD53RZ	2025-06-02	Lyft	travel	51.45	USD	paypal	6714086308777	2362958769587	debit
181	O2G5TE8JXRTSO7	2024-08-01	Lyft	travel	73.82	USD	apple pay	3635635279133	5005797434049	debit
188	8UCX92APT97OX8	2024-11-08	IRCTC	travel	3429.69	INR	netbanking	6301462696703	7715042288682	debit
189	H1ZDUTU2BYVAY4	2025-06-24	Uber	travel	4407.47	INR	cash	6301462696703	451707923874	debit
191	5708UFJIWQ6MTD	2025-01-04	Uber	travel	1363.3	INR	cash	1478276631136	1188575632362	debit

3.**Total amount spent on groceries using netbanking in INR:**
   - Sum of amounts where `category` is "groceries", `payment_mode` is "netbanking", and `currency` is "INR".
   - Total: 4,386.24 + 476.05 + 2,899.52 + 4,396.34 + 3,572.75 + 2,953.14 = 18,684.04 INR.



transactions_for_prompt_gen

	transaction_id	date	merchant	category	amount	currency	payment_mode	receiver_acc_no	sender_acc_no	type
11	I2ZZ2M16QAGTHH	2025-02-28	D-Mart	groceries	4386.24	INR	netbanking	6249501219399	5736305731539	debit
73	X2QA272I0KATQC	2024-12-22	Reliance Fresh	groceries	1875.38	INR	netbanking	6249501219399	5997695932956	debit
88	RN7P8FO87D7E3A	2024-09-02	Spencers	groceries	2953.14	INR	netbanking	5040614351036	6791651547245	debit
167	WZMZDC2QDM4KVP	2025-03-01	Spencers	groceries	419.37	INR	netbanking	7795839006956	1139775773716	debit
172	FQJBO4UHUSM366	2024-08-07	Spencers	groceries	4396.34	INR	netbanking	6301462696703	6387471978442	debit

While the model performs arithmetic operations correctly over the retrieved data, it does not always retrieve **all matching transactions**, leading to partial totals. This highlights the dependence on the retrieval stage's ability to fetch complete context.

Query Type & Example	Observed Model Output
Statistical: Total amount credited from job-related transactions in USD.	Model summed relevant transactions to **31,624.83 USD**. Calculation is arithmetically correct, but may not always retrieve all matching rows from context.
Statistical: Average transaction amount for travel expenses.	Calculated total **30,404.92 USD** across **24 transactions**, giving an average of **1,266.87**. Shows partial data grounding with potential misses in edge cases.
Statistical: Total spent on groceries via netbanking in INR.	Reported **18,684.04 INR**, accurately summing known transactions. Still depends on what retrieval stage fetched.
Hallucination check: December 2024 transactions.	Retrieved **7 major transactions totaling 17,633.77 INR**, missed some minor or edge rows, indicating retrieval sensitivity.

Future work

Future work will focus on:

- Adding **metadata filtering** (by month, category) before vector retrieval to guarantee all matching transactions are considered
- Testing with **hybrid retrieval (dense + sparse keyword filters)** to capture missed entries.
- Running more **counterfactual tests**, e.g., queries for months with no grocery transactions, to better gauge false positives or speculative reasoning.
- Exploring lightweight **fine-tuning or LoRA adapters** on synthetic financial QA data to better specialize the LLM on aggregation queries.

This work lays the foundation for trusted, domain-specific assistants that preserve user privacy without compromising on answer relevance or quality.

Additionally, incorporating audit logs to track which document chunks contributed to each answer would further enhance trust and explainability

5: Bibliography

The following are referred journals from the preliminary literature review.

- Lewis et al., 2020 – RAG: Retrieval-Augmented Generation introduced combining dense vector retrieval with transformers for grounded question answering. Link
- Meta's LLaMA, Mistral, and Ollama have enabled local deployment of large LLMs without cloud dependencies, addressing growing concerns about privacy. Link
- Academic tools like **FAISS** and **sentence-transformers** are proven components in semantic search pipelines. Link

Appendix

This section includes sample synthetic data, example prompts used in evaluation, select code excerpts, and additional observations supporting the experimental results discussed in the main report.

Sample synthetic bank statement data

1	transaction_id	date	merchant	category	amount	currency	payment_mode	receiver_acc_no	sender_acc_no	type
2	TT40OHQY68VY0I	2024-07-01	Company Payroll	job	44442.52	INR	netbanking	6255837252251	7713815500479	credit
3	SUW28E4DT04EZF	2024-08-06	Company Payroll	job	54431.35	INR	netbanking	6255837252251	7197144893682	credit
4	WFGWO9X5W74ADN	2024-09-03	Company Payroll	job	22743.5	INR	netbanking	6255837252251	8001357967908	credit
5	CANH8NH486TOYE	2024-10-03	Company Payroll	job	39708.68	INR	netbanking	6255837252251	6763272083341	credit
6	IB52Q0KGW451R9	2024-11-06	Company Payroll	job	32453.7	INR	netbanking	6255837252251	3469035989871	credit
7	MMZFYRQR25T12X	2024-12-03	Company Payroll	job	37534.25	INR	netbanking	6255837252251	9825898175440	credit
8	5KYNBDLZ8VQ1H6	2025-01-07	Company Payroll	job	33059.32	INR	netbanking	6255837252251	6657038038402	credit
9	A7DQ8ZKG6IYVXM	2025-02-01	Company Payroll	job	55501.26	INR	netbanking	6255837252251	8754334060777	credit
10	T4RB34K6Z1D383	2025-03-06	Company Payroll	job	63480.89	INR	netbanking	6255837252251	1880112392372	credit
11	RK83D0OAMDWHZS	2025-04-04	Company Payroll	job	84504.57	INR	netbanking	6255837252251	5336660980990	credit
12	MBS9LTTFVSKCVR	2025-05-01	Company Payroll	job	24838.42	INR	netbanking	6255837252251	0570613843685	credit
13	07SSARTU734IRN	2025-06-02	Company Payroll	job	32350.17	INR	netbanking	6255837252251	0285611227661	credit
14	A2IIKBT163INGV	2024-08-10	D-Mart	groceries	579.29	INR	netbanking	6255837252251	2400931895771	debit
15	54BWW2EIZB5AOG	2024-09-21	Flipkart	online_shopping	3808.82	INR	cash	6255837252251	3050135398074	debit
16	HLW2LSN182EDDJ	2024-09-20	Spencers	groceries	4486.39	INR	UPI	6255837252251	9117936694549	debit
17	45ZJZZIOPCLNVB	2024-10-31	Mobile Recharge	utilities	369.72	INR	UPI	6255837252251	8907459243521	debit
18	4WBOU5AXJYN4UX	2025-03-18	Mobile Recharge	utilities	2762.37	INR	cash	6255837252251	1236964625587	debit
19	SNOAOUKULY8ZQQ	2025-03-05	Petrol Bunk	utilities	1981.65	INR	card	6255837252251	9179673604912	debit
20	AGI3CM2H27QWLU	2025-01-08	Amazon India	online_shopping	2692.79	INR	card	6255837252251	0947117959816	debit
21	E9EB0UH84ABCYR	2025-06-16	Groww	stocks	11291.84	INR	netbanking	6255837252251	7091290793310	credit

1	transaction_id	date	merchant	category	amount	currency	payment_mode	receiver_acc_no	sender_acc_no	type
2	9NIFF2RYSRWUYF	2024-07-03	Company Payroll	job	1895.87	USD	netbanking	5553014498758	7970984708208	credit
3	TT4NOS5N1XJHKU	2024-08-01	Company Payroll	job	582.07	USD	netbanking	5553014498758	8818103331077	credit
4	3OPPN2CP8Y7DRQ	2024-09-02	Company Payroll	job	1664.81	USD	netbanking	5553014498758	4940126576522	credit
5	8XVBRTATMX0F1V	2024-10-04	Company Payroll	job	522.9	USD	netbanking	5553014498758	9030105358254	credit
6	RMOH1ZZ23KO9UL	2024-11-01	Company Payroll	job	1841.69	USD	netbanking	5553014498758	5400230252314	credit
7	N72YHFHATCPM80	2024-12-03	Company Payroll	job	1648.95	USD	netbanking	5553014498758	8039450980571	credit
8	6ZADP9FBSHYZN8	2025-01-01	Company Payroll	job	639.18	USD	netbanking	5553014498758	5940476061135	credit
9	4DEDQM0POF8HX2	2025-02-07	Company Payroll	job	2520.27	USD	netbanking	5553014498758	8816758171996	credit
10	Y6ISW6QFHUQKR8	2025-03-05	Company Payroll	job	2321.31	USD	netbanking	5553014498758	6365064083936	credit
11	8EMAAAY3QONBPV	2025-04-03	Company Payroll	job	1703.44	USD	netbanking	5553014498758	0157859042269	credit
12	J3O29IUP0V60B6	2025-05-05	Company Payroll	job	2185.05	USD	netbanking	5553014498758	5374762662877	credit
13	YSWJAFQ1NOOAKY	2025-06-02	Company Payroll	job	851.37	USD	netbanking	5553014498758	4656453794224	credit
14	5J6S9DAEXMTXM9	2024-12-21	Starbucks	others	99.7	USD	cash	5553014498758	3765504311141	debit
15	PIXIZJ7OGIDUFD	2024-11-30	BlackRock	funds	1051.62	USD	netbanking	5553014498758	2298597833993	credit
16	U02VYQDICTUMA8	2025-05-12	Starbucks	others	59.01	USD	card	5553014498758	5524503814947	debit
17	L4QKBQDIPEQIDS	2024-12-22	Lyft	travel	80.43	USD	card	5553014498758	8758402838645	debit
18	O8GK9X3AIW2UZ0	2024-10-30	Starbucks	others	80.44	USD	netbanking	5553014498758	2187191425592	debit
19	MTM2RGB92AELDI	2024-07-17	Starbucks	others	91.29	USD	netbanking	5553014498758	9085303548329	debit
20	UZRDSL5TKL00OM	2024-10-15	Comcast	utilities	21.58	USD	card	5553014498758	4020900560960	debit

Prompt Examples Used for Evaluation

1. Calculate the total amount credited to the account from all job-related transactions in USD.
2. Determine the average transaction amount for all travel-related expenses across all payment modes.
3. What is the total amount spent on groceries using netbanking in INR?
4. Calculate the average amount of all online shopping transactions made through card payments.
5. Identify the trend in spending on utilities over the months and describe any patterns observed.
6. What is the total amount credited from all funds-related transactions in INR?
7. Calculate the percentage of total transactions that fall under the category of travel.
8. Determine the average amount spent on groceries per transaction in USD.
9. Summarize the monthly trend of debit transactions made through the UPI payment mode.
10. What is the total amount spent on online shopping using Paypal?
11. Calculate the total amount spent on others category using card payments in INR.
12. What is the average amount of funds-related transactions made via netbanking?
13. Determine the total number of transactions related to job income.
14. Calculate the average transaction amount for all debit transactions across all categories.
15. What is the total amount spent on utilities using cash payments in INR?
16. Determine the month with the highest total expenditure on groceries.
17. Calculate the average amount of credit transactions in the stocks category.
18. What is the total amount spent on travel using apple pay?

Environment or Config Summary

- Hardware: MacBook M1 / Ubuntu VM
- Libraries: sentence-transformers, FAISS, Ollama, Pandas, Fakr
- LLM Model: Mistral 7B via Ollama, OpenAI
- Python